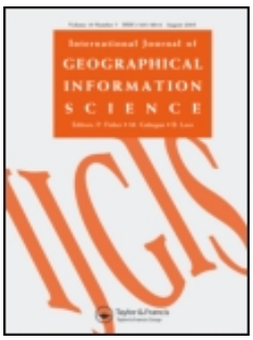

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Extracting geographic features from the Internet to automatically build detailed regional gazetteers

Daniel W. Goldberg , John P. Wilson & Craig A. Knoblock

To cite this article: Daniel W. Goldberg , John P. Wilson & Craig A. Knoblock (2009) Extracting geographic features from the Internet to automatically build detailed regional gazetteers, International Journal of Geographical Information Science, 23:1, 93-128

To link to this article: http://dx.doi.org/10.1080/13658810802577262

Published online: 06 Apr 2009.

International Journal of Geographical Information Science
Vol. 23, No. 1, January 2009, 93–128

Taylor & Francis
Taylor & Francis Group

Research Article

Extracting geographic features from the Internet to automatically build detailed regional gazetteers

DANIEL W. GOLDBERG*†, JOHN P. WILSON‡ and CRAIG A. KNOBLOCK§
†Department of Computer Science, University of Southern California, Kaprielian Hall 444, Los Angeles CA 90089-0255, USA
‡Department of Geography, University of Southern California, Kaprielian Hall 444, Los Angeles CA 90089-0255, USA
§Department of Computer Science, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292, USA

(*Received April 2006; in final form May 2008*)

The utility of every imaginable application which incorporates a gazetteer hinges on the simple fact that the resulting system will only be as useful, complete, or accurate as the underlying gazetteer itself. A major issue confronting gazetteers utilized in systems today is that they are not complete and measures of their accuracy are largely unknown. In this paper we describe a methodology which addresses this problem by automatically generating highly complete and detailed regional gazetteers from Internet sources. We utilize information extraction and integration techniques to automatically obtain geographic features and associated footprints and feature types from freely and widely available online data which could be applied to create a gazetteer for nearly any area. We discuss the distinguishing characteristics of the generated gazetteer and extend previous work to define measures which can be used to assess the completeness and accuracy of gazetteers. Using these measures, the generated gazetteer is evaluated against the Alexandria Digital Library Gazetteer and the Los Angeles Comprehensive Bibliographic Database. Our results indicate that a gazetteer created by our methods will be at least as complete as any gazetteer currently available for certain feature classes, while falling short in others. We conclude by offering suggestions to address these shortcomings.

Keywords: Gazetteers; Geographic information extraction

1. Introduction

As defined by Hill (2000) a gazetteer is a list of geographic features consisting of three main parts; the name (or toponym), type, and footprint. The primary role of a gazetteer is to translate the vernacular into the scientific, allowing users to seamlessly move back and forth between a world where entities are spoken about in human language and one where entities are scientifically grounded within concrete geographic footprints and well known types (Goodchild 1999). This simple gazetteer definition provides the foundation for a wide variety of tools and applications in many diverse fields including the digital library, natural language processing and

*Corresponding author. Email: dwgoldbe@usc.edu

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810802577262

linguistics, and intelligence communities. Each of these applications requires a different level of accuracy and completeness.

Applications requiring very high detail are increasing in number every day, especially with the explosion of web-based services on the Internet. However, there are very few current gazetteers that can be described as highly complete or accurate. The reason for this is that it is simply far too time-consuming to manually create gazetteers at a very fine level of granularity (see Hill 2000, Wilson *et al.* 2004 for a discussion of the problems and challenges of creating footprints, for example). Some of these problems disappear when computers are used and can spend countless hours submerged in this task without complaint, and there is no reason to prevent a gazetteer from being as detailed as possible if the information is available. It is always possible to abstract lower levels of granularity from a highly detailed gazetteer, but it is seldom possible to go in the other direction, creating a highly detailed gazetteer from a less detailed one.

With the vast amount of geographic information available on the Internet, one can utilize information extraction and information integration techniques to automatically obtain geographic features and associated footprints and feature types from freely available online data to automatically generate gazetteers. While the information extraction tools that can facilitate the automatic creation of gazetteers from online sources are themselves mature and capable of being rapidly deployed to extract and integrate data from any number of online sources (Knoblock *et al.* 2000), their suitability for the task has yet to be seen. This paper will investigate and discuss the distinguishing characteristics of a gazetteer generated using these techniques and extends previous work by defining measures to assess the completeness and accuracy of this and other gazetteers.

The remainder of the paper is organized as follows. First, we argue for the necessity of highly detailed gazetteers and discuss measurements of accuracy. We then outline our data sources and the methodology used to automatically generate a gazetteer, and report a detailed analysis of the results, followed by an examination of related work. We conclude by highlighting some defining characteristics and shortcomings of gazetteers created using our methods and offer suggestions for future work.

2. Highly detailed gazetteers

The utility of every imaginable application which incorporates a gazetteer hinges on the simple fact that the resulting system will only be as useful, complete, or accurate as the underlying gazetteer itself. The more detailed the gazetteer, the better results a user will be able to expect from the system. Two fields which rely heavily on gazetteers where this fact can be clearly seen are feature identification in imagery and named entity recognition (NER).

2.1 *Feature identification*

High resolution satellite imagery, typically 1 m or less, is now readily available for most of the USA as well as other parts of the world (e.g. Microsoft Corporation 2008a, b, Google, Inc. 2008, Yahoo!, Inc. 2008). It is often useful for applications to be able to automatically identify features in a satellite image or aerial photograph to aid in tasks such as conflation of data from multiple sources (Chen *et al.* 2004), or for the simple annotation of the features on the image (Barclay *et al.* 2000, Chen

et al. 2003). To accomplish this, one could consult a gazetteer to obtain every feature within the scope of the area of the image and overlay the footprints from the gazetteer on top of the image. For this to produce an accurate result, the gazetteer would need to have complete coverage for the area, meaning that every feature in the image would have to be represented in the gazetteer, otherwise there would be certain features not present or without information. Additionally, the footprints for each of the features in the gazetteer would need to be correct, or the user of this system could mistake one feature for another, sometimes causing catastrophic results.

2.1.1 A matter of scale. At this point, with the feature identification application in mind, we should take a step back and ask ourselves a few fundamental questions regarding the 'scale' or 'resolution' of a gazetteer. What is the appropriate level of granularity for features in a gazetteer? How far should we break the world down into pieces such that each can be individually identified? Should every inch of the globe be a feature, since each can be uniquely identified by its geographic coordinates down to any arbitrary level of accuracy? Should gazetteer features be only large scale features like countries or cities? Are they only non-manmade features like mountains with a few specific man-made features thrown in (e.g. The White House)? Affirmative answers to the last two questions represent the status-quo in conceptual thinking about gazetteers in that they are typically regarded as 'low-resolution' data structures containing large scale features.

However, if one accepts the premise of using a gazetteer for feature identification (as in the previous section), it becomes obvious that 'low-resolution' will not suffice. If one's task were to identify every structure in an image in terms of its address or name (if applicable, e.g. One Wilshire, or the Roosevelt Hotel) with its classification (i.e. type, e.g. commercial, residential) and footprint in a satellite image, a gazetteer of what would, presently, be considered 'ultra-high-resolution' features would be required (parcels with addresses, building footprints with addresses or names, etc.). In this paper, we will introduce the concept that there need not be a minimum or maximum scale associated with features in a gazetteer. We argue that anything which can be uniquely identified or addressed (has a name), can be classified (has a type), and takes up geographic space (has a footprint) should be an acceptable feature for inclusion in a gazetteer. As previously mentioned, one does not always know the uses for which a gazetteer will be employed after it is created, and with information storage now cheap, there is no reason to preclude the existence or creation or these types of 'ultra-high-resolution' gazetteers.

Now of course, we realize that manhole covers and sewer grates fulfill the requirements we have just laid out. To test the feasibility of our definition let us explore the question: Should a list of manhole covers or sewer grates comprise a gazetteer? To someone interested in accessing information in a digital library context, of course not. Maintaining these lists offers no benefit for the lower-resolution queries (e.g. named places) which they would issue to have indirectly georeferenced. However, for emergency response personnel trying to determine the potential spread of a toxic liquid after a tanker truck overturns while traveling on a freeway, this would be an invaluable resource. They would be interested in the spatial footprints of specific features with specific feature types (manholes and sewer grates) which are within the area of interest that could potentially cause the liquids to seep into the sewage system, in order to potentially evacuate the proper people in the path of the toxins. They would want to be able to spatially query a typed list of

utility features (of which manhole covers and sewer grates would be a subclass) and have features with footprints returned.

It should be obvious at this point that these lists do in fact constitute gazetteers. While the toponym axis may not be as important as the spatial footprint or typographic axis because the personnel do not need to address them by name [and they are most likely named as codes in a linear addressing system (Fonda-Bonardi 1994) as are most public utility infrastructure, e.g., overpasses, tunnels] the three axes of the gazetteer are indeed still there. What the gazetteer research community will need to accept [and what some already are (Agouris *et al.* 2000)] is that a gazetteer is as much an outright spatial data model as it is a typographic classification system, and that this role carries as much weight as any other the gazetteer performs.

While these types of ultra-high-resolution databases surely exist at utility companies around the world, they are seldom referred to as gazetteers. This, however, is not the case across the board; some utility infrastructures are indeed included in gazetteers already. In fact, the most often cited gazetteer in research, the Alexandria Digital Library (ADL) Gazetteer (Alexandria Digital Library 2008), contains utility features along these same lines since it includes radio transmission towers. Being more high-profile and far fewer in number than sewer grates, it is understandable why they are presently included and sewer grates are not [since this state of affairs is consistent with Hill's satisficing condition (Hill 2000)].

Further, we can see the true cause of our present situation. Simply put, it is a matter of scale (Smith and Mark 1998, Mark *et al.* 2001). Geographic features are not considered gazetteer features until an application which needs them (i.e. that uses data at the same scale as the feature) is developed. Thus, in the domain of non-typical geographic features (e.g. sewer grates), once the designers of ultrahigh-resolution geospatial applications finally realize that the data they are storing are in fact gazetteer data (i.e. that they are geographic features which are named, typed, and have locations), and the structures and access methods they utilize to work with them (i.e. querying by name, type, or footprint), do in fact constitute gazetteers in every sense of the word, the sewer grate's time to shine as a true geographic and gazetteer feature will have arrived. Undoubtedly these types of highly-specialized gazetteers (and accompanying feature type hierarchies) will raise some eyebrows, but distant cousins in the form of highly-specialized gazetteers already exist and are widely accepted as such [e.g., historical and religious sites (Electronic Cultural Atlas Initiative 2008, Berman 2003, 2004)].

Further examples of the scale-dependent validity of gazetteer feature types can be seen when investigating the classic gazetteer feature type: bodies of water. For example, the Santa Monica Bay and its parent feature, the Pacific Ocean, which border Los Angeles (LA), would certainly be considered valid features in most gazetteers, and are low-resolution (i.e. cover a large area). Now, typically when it rains in LA, raw sewage and trash flow off the streets, down into the sewage system, and out into Santa Monica Bay making water conditions extremely harmful, with organizations such as Heal the Bay and Surfline recommending people not enter the water after examining samples from specific sites along the bay (Heal the Bay 2008, Surfline/Wavetrak, Inc. 2008). However, the local knowledge possessed by these organizations allows them to further divide the larger Santa Monica Bay gazetteer feature into smaller features describing each of the particular (non-officially named, yet extremely well known) wave breaks where people surf, e.g. County Line,

Topanga Point, and El Porto. These divisions are extremely important to people interested in determining where it is safe to surf because they indicate at a high-resolution where one may enter the water, which would not have been possible by simply using the lower-resolution term Santa Monica Bay. At the scale of the 'application' being used (e.g. surfing spot determination at the local level), these named, typed (wave breaks), and spatially located features should again be considered valid gazetteer features, while for a tsunami-warning system utilizing a gazetteer of water bodies, they most likely would not be because they are too high-resolution.

2.2 *NER*

Perhaps even more reliant on gazetteers are NER techniques which play a critical role in many domains and applications such as webpage classification (Amitay *et al.* 2004, Martins *et al.* 2005a), toponym resolution and disambiguation (Leidner 2004a, Li *et al.* 2002, Smith and Crane 2001), and some historical collections and digital geolibraries. In these, all of the holdings are georeferenced, meaning that they have an associated geographic footprint, allowing a user to search for holdings relevant to a particular geographic region in a query (Janee *et al.* 2004) such as 'all maps, news articles, and aerial photos about the region within a certain geographic bounding box'. The process of manually reading every non-geographic text and assigning a geographic footprint is not practical, so NER techniques are typically used to identify geographic feature names in the text (Beaman *et al.* 2004, Berman 2004, Reid 2003, Witten *et al.* 2004) which can then be looked up in a gazetteer to obtain a geographic footprint.

It has been shown that NER systems work well for identifying all traditional types of entities (people, organizations, and so on), except for geographic features (Mikheev *et al.* 1999). As such, NER systems generally incorporate a gazetteer to help determine if a word is a geographic entity, and if it is, what its geographic footprint should be. For a NER system to accurately identify a geographic name in a text, it usually requires the existence of the named geographic feature appearing in the text to be in the gazetteer. Depending on the type of text document being examined, the required gazetteer may need varying levels of granularity and completeness (e.g. the scale-dependency previously noted). For instance, text documents such as international news articles could possibly be processed using a gazetteer with a lower degree of completeness, containing only countries and major city names. Alternatively, a local newspaper might require a very high degree of completeness, including the names of small cities, communities, neighborhoods, local business, and even apartment complexes. Processing ancient and foreign texts requires a completely different gazetteer (Chavez 2000, Berman 2004, Buckland and Lancaster 2004), as would textual descriptions of locations (Wieczorek *et al.* 2004). The wide breadth of possible topics and deep specificity of features types and cultural differences have historically been hindrances to the creation of a single gazetteer that accounts for all historical place-names, but projects and consortiums such as ECAI are attempting to address this and other issues (Electronic Cultural Atlas Initiative 2008).

Another primary role of the digital geolibrary which is highly reliant on a gazetteer is indirect georeferencing (Hill and Zheng 1999), whereby a user of the library is able to issue a search for the name of a place, and that name is then translated into a geographic footprint and a geographic search is performed on the

library's holdings. For this to succeed, the gazetteer must contain the term entered by the user or it will be impossible to translate it into a geographic location that can be used as a spatial query on the georeferenced holdings of the library, unless some form of translation to an existing name occurs. Even though relevant data may exist in the collections, nothing will be returned to the user if the term they entered does not exist in the gazetteer. The more detailed the underlying gazetteer is, the better the chance that the search term entered by the user will be successfully handled by the system. However, higher detail does not necessarily mean that an input query will be handled better; one must still be aware of the quality of the reference data features and how this may affect the rate of false positives, or imprecise results.

3. Accuracy

A major issue confronting gazetteers utilized in systems today is that they are not complete, and measures of their accuracy are largely unknown. Furthermore, there is a fundamental lack of consensus as to what the terms 'completeness' and 'accuracy' should refer to when speaking in terms of gazetteers. Is accuracy referring the features themselves on an atomic measure, e.g., the spatial accuracy of the feature footprints in terms of distance from true location on the ground, the accuracy of the feature name in terms of temporal or cultural validity, or the accuracy of the feature type in terms of specificity and granularity? Or, does accuracy refer to measures of the gazetteer as a whole, in a holistic sense, describing aggregate or general characteristics of the three axes? Measures of completeness are similarly subjective and applicable at both the atomic and holistic scales, e.g., the completeness of a single spatial footprint (with a polygon being more accurate than a centroid) versus aggregate counts of the number of features with particular types of footprints, the completeness of a single feature's name (with more aliases or historical names being seen as more complete) versus aggregate counts of the prevalence of multiple names, or the completeness of the feature type (with the association of more types being more accurate) versus a simple measure of the sheer number of features types in the gazetteer as a whole.

Any one of these measures (plus many more) can be used to describe the quality of the three axes of both features themselves as well as the gazetteer in its entirety. As discussed previously, these measures will be highly dependent on the applications for which they are being used, with particular usages necessitating 'completeness' and 'accuracy' to mean different things. Different applications will require different levels of granularity as to what types of features are represented, different levels of completeness in terms of the feature names that are included, and different levels of detail for feature footprints, all being valid if they are capable of helping individuals make decisions on the appropriateness of a particular gazetteer's application to a particular task.

As noted by Goodchild (1999) and again by Leidner (2004b), it does not make sense to talk about the accuracy of a gazetteer without a notion of granularity or scale, in this case referring to the level of detail expected from a gazetteer. Depending on the consumer of the gazetteer, coarse features defined down to the level of country or city may be sufficient. Other gazetteers may require very fine grained features such as schools and police stations. It is unfair to claim that a gazetteer is incomplete because it only contains feature classes a, b, and c, while not containing anything of the more detailed types x, y, and z because the creators of the gazetteer may not have intended it for use in the highly detailed domains of x, y, and

z. Cities represented as points in one gazetteer may be sufficient for the application for which the gazetteer was constructed, while being far too coarse for other applications. Essentially, with current standards and practices, no one single gazetteer can ever be considered complete for all tasks because each gazetteer is designed and created with some pre-determined task in mind which may or may not be applicable to others.

Both of the most well known gazetteer standards, the ADL Content Standard (Alexandria Digital Library 2006) and the Open Geospatial Consortium (OGC) Gazetteer Service Profile of a Web Feature Service (Open Geospatial Consortium 2002) (WFS-G), allow individual features to have measures of accuracy associated with them, although neither of these includes a measure of completeness, correctness, or accuracy for the gazetteer as a whole.

In their ongoing research, Doerr and Papagelis (2007) attempt to measure the completeness and correctness of gazetteers based on the number of successes and failures in place-name lookups. Leidner (2004b) introduces seven measures which can be used to describe a gazetteer including availability, scope, completeness, correctness/precision, granularity, balance, and richness of annotation. We will extend this work to include additional measures for the completeness and accuracy of the gazetteer based on the physical coverage of the features on the ground.

4. A parcel-level gazetteer

What we have attempted to do throughout the preceding section is show with the use of specific examples that the scale of what is an acceptable gazetteer feature type will be dependent on the application for which it is employed. We feel that the gazetteer research community should celebrate the creation of ultra-high-resolution gazetteers, because the more people and applications which utilize them, the greater the need (and funding) for research into the core areas of gazetteer development, benefiting every user from low-resolution to ultra-high-resolution. Our discussion will next turn to a specific class of high-resolution features: the land parcel. We will first illustrate the critical need for these types of gazetteers, and then discuss the challenges in obtaining them, motivating the Internet-extraction based methodology presented in the remainder of the paper.

4.1 *The need*

In urban areas, we can confidently state that the predominant geographic feature types present (i.e. those that occupy a majority of the space) would be man-made structures (i.e. buildings of varying types and usages), roads, and green spaces (i.e. parks). Geospatial applications requiring high-resolution spatial models of these urban environments (particularly buildings) are ubiquitous, and the information required to be maintained for the geographic features are most commonly an identifier (name), a type, and a footprint, i.e., valid gazetteer data according to our definition.

Thinking in terms of the spatial models used, these applications require the landscape to be broken down to the scale of individual personal activity. Thus, at a minimum, the scale of the geographic features present must be the parcel, with actual building footprints highly desirable. However, these actual physical building footprints are often extremely difficult to obtain, so parcel data are commonly used in their place (Henson and Goulias 2006). One example of a research field that

utilizes this level of detail in urban areas is the transportation research community, given their long tradition of using micro-scale spatial models (including down to the resolution of parcels) to create, analyze, and predict transportation models and perform micro-simulations (for a review, see Henson and Goulias 2006).

In particular, the fundamental role of the transportation research community in 'homeland security' research and practice has identified the urgent need for parcel databases with activity-type associations (i.e. a classification for the structure) to prepare for and respond to unexpected events (Henson and Goulias 2006). Leading researchers have even gone so far as to state 'For complete coverage, the ideal and most detailed geographic unit is a parcel of land' (Henson and Goulias 2006: 10). For these and the other gazetteer-consuming applications discussed earlier (feature identification and NER) one should strive to have the most complete and accurate underlying model possible. If parcels are the fundamental building blocks of your geography, then a gazetteer of parcels would fill this need perfectly.

Let us consider next what information must be maintained about each feature. Obviously, a spatial footprint is required, otherwise there would be no way to ground the features in a geographic domain, making them useless for any type of spatial analysis. Typographically, one would want the features classified in terms of their usage. Minimally, distinctions should be made between residential (where people live) and commercial (where people work) locations. Further distinctions could be made within each category such as single-and multi-family in the residential branch, and customer oriented (i.e. where customers actually frequent the establishment, e.g. restaurants, car mechanics, and banks) and manufacturing (i.e. where customers usually do not go, e.g. textile factories). These distinctions would be necessary in the transportation domain for micro-simulations modeling flows between locations throughout the workday. For example, from single-family residential (home) to manufacturing (work) in the morning, to customer-oriented in the afternoon (lunch), back to work, to customer-oriented after work (dry cleaning), and back home in the evening. Finally, toponymically, one needs to maintain some identifier. This could be the postal address of the property, the name of the owner, the name of the business, etc. Depending on the application one might be preferable to another, but the availability of information will ultimately drive what is chosen. In some areas, owner names for parcels might be available while in others they are not. At a minimum, we will consider a postal address sufficient as a proper toponym, with additional information included as available. We feel this assumption is valid because the minimum granularity we will consider for a feature in this paper is the parcel, for which most municipalities require a valid address. In terms of completeness and accuracy, one should strive to have correct descriptive information (accuracy) about every parcel within the area of interest (completeness).

One major hurdle standing in the way of the creation of gazetteers of parcel-level or addressable features is personal privacy, a fact long known in the health research community. In particular, the requirements for health studies working with human subjects to obtain the approval of Institutional Review Boards (IRBs) and researchers to acquire training for human subject research has become (required) standard practice at research institutions worldwide (being federally mandated in the US). In the case of health research, these research requirements are intended to protect the rights, privacy, and confidentiality of the individuals whose personal data are being worked with.

These rules ensure that it is forbidden for any type of personally identifiable information to be released at any point prior, during, or after any part of the study. But how should these requirements translate to other fields such as transportation or homeland security? Is the creation of a parcel-level or addressable feature gazetteer violating the rights of the parcel owners? Should the businesses contained with their names, types, and building footprints be afforded the same protection as those individuals in a health study? For example, no one doubts that health studies monitoring outbreaks of potentially disastrous diseases fall under the umbrella of homeland security, and, as a health study, should be subject to the very same IRB scrutiny just described. However, should a study attempting to create a named, typed, and spatially oriented model of the buildings within an area be subject to the same rules? There are, presently, no clear answers to these questions. On the one hand, some key personal identifiable information (owner names and addresses) will be maintained in the database, and its creation should therefore be forbidden. On the other hand (as we will see throughout the remainder of the paper), this information already exists in readily available formats, and is presented in some cases as 'public information' by the very same government agencies tasked with serving and protecting the public.

We will not attempt to solve this dilemma through the research presented in this paper. This issue is far too contentious with repercussions up and down the data food chain (both public and private sector) that would and should have effects on data and privacy policies everywhere (for a recent review of privacy laws and policies spanning numerous jurisdictions, see Cho 2007). In this paper, we will employ publicly available data sources (which anyone connected to the Internet can access), to test a methodology for integrating these data in novel and useful ways. While the consolidation of these data may contain personal information, this very same personal information is provided as 'public information' by the government sources from which it is derived. Its release and subsequent discussions of its appropriateness should be a larger policy issue directed to the government agency releasing it. In the same vein that open-source software development leads to the production of ultimately better software through open exchanges of ideas and cross-institution testing and development, we believe that by highlighting how the information we obtain and exploit can be used, we will ultimately provide definitive evidence of its usefulness and accuracy for particular applications as well as potentially broader policy discussions.

4.2 *The challenges*

As pointed out by Lawson (2005) and the TRANSIMS experiments performed by the Los Alamos National Laboratory described in Henson and Goulias (2006), administrative databases containing these types of information, while crucially important, are sporadically available and fusing them together presents an arduous task because there presently exists no national-scale parcel database [although the development of one is being investigated by the Federal Geographic Data Committee (FGDC) Subcommittee for Cadastral Data (Stage and von Meyer 2005)], and bureaucratic hurdles will need to be overcome to produce such a database (Dawes *et al.* 2006).

For instance, the Attorney General of the State of California wrote a legal opinion in 2004 stating that 'A copy of parcel boundary map data maintained in an electronic format by a county assessor must be furnished 'promptly' upon request of

a member of the public' (Lockyer 2005: 2), yet it has taken years for this to be implemented [e.g., LA County in 2006 (Auerbach 2006)], and currently not all counties are providing it, or providing data files with all necessary attributes. As an example, parcel data acquired from the LA County Assessor's Office in April 2006 does not include any attributes other than spatial geometry and assessor identification number (AIN). Without the address or land-use type attributes, the usefulness of these data in homeland security applications is severely limited.

What we can see through the assessments of the FGDC (Stage and von Meyer 2005) and the practical experience of obtaining data from the LA County Assessor's Office is that a significant amount of time will be required before official parcel data sets are readily available for researchers to use. However, this does not lessen the urgent need for these types of data sources in the research fields that require them (i.e. the transportation and homeland security fields previously discussed). With this in mind, the remainder of this paper presents and tests a proof-of-concept methodology that can be applied to generate a gazetteer of parcel-level gazetteer of addressable features that would be extremely useful for these applications.

4.3 *An alternative*

In cases when a parcel-level gazetteer of addressable features is not available or not of sufficient quality, yet a research project calls for it, how should one proceed? Should the potential research be tabled until the data become available? We argue that it should not be; that the information required exists and is readily available. The methodology developed to perform the experiments in this paper provide a proof-of-concept that this information can be gathered and integrated to construct a proper gazetteer for use in the high-resolution applications previously mentioned. We will show that as an alternative option to using a parcel database directly, one can utilize this information from online sources containing descriptive information about parcel-level addressable geographic features to generate a high-resolution gazetteer. Further, and perhaps most importantly, an analysis will be presented to determine the quality of the resulting data in relation to the alternatives which are presently available for use in research.

At this point, we should be asking ourselves questions such as what data sources and/or tools are required to derive the necessary information, and once compiled, what quality will the resulting data be?

5. Automatic gazetteer creation

Our method for automatically generating a gazetteer relies heavily on techniques from the fields of information extraction and information integration, subfields of computer science focused on gathering and integrating diverse sets of data from disparate sources, especially information available from Internet sources. These topics have been the focus of considerable research in recent years as the availability of information on the Internet has exploded. The semi-supervised learning techniques pioneered in this field enable the rapid creation of agents by simply marking up examples of data to be extracted on relatively few training samples (for an overview and taxonomy of major data extraction tools, see Laender *et al.* 2002). These tools allow one to quickly wrap online sources of semi-structured text (for example web pages) and create agents that can then be treated as structured information sources or online databases. Most modern software systems which

support this type of agent creation and information extraction support the extraction of data from web pages in any language and any encoding, allowing one to extract information from and about web pages for any part of the world. We can utilize these information extraction and information integration techniques to extract geographic features from Internet sources to rapidly produce very detailed regional gazetteers.

Throughout this research we used the AgentBuilder software created by Fetch Technologies (Fetch Technologies, Inc. 2008) to build our agents. The agent creation process implemented with this tool consists of modeling the navigation structure of the site, as well as creating a schema of information to be extracted and marking up training examples consistent with the schema. The software then uses the training examples to learn the general rules necessary to extract the data for each piece of the schema. The agent produced by the software contains these extraction rules and the navigation structure required to move throughout the pages of the site to accomplish the extraction goal. The output of the agent can go directly to a database or be streamed back in XML. Full details on the agent building process using the AgentBuilder software are available in Beach *et al.* (2004). This approach is commonplace for applications which use information extraction tools for gathering web data, as we have done in our previous work (Bakshi *et al.* 2004).

It is worth noting that in the foreseeable future, as true GIS Web Services become more commonplace, this requirement for the use of agents to obtain data may diminish as application programmer interfaces (APIs) are developed and exposed for data-rich and intensive applications such as ours.

As noted by Berman (2004), the sources selected for creating a gazetteer (in our case those chosen for extracting geographic features) have an enormous impact on the resulting gazetteer. Obviously, the gazetteer will only contain features that are present in the source, and choosing a more complete source will result in a more complete gazetteer. In addition, the feature types added to the gazetteer will be limited to those in the source. For this reason it is useful to contrast our approach, where we identify sources and gather data, from one that simply crawls the web and obtains or indexes any data it finds (e.g. Google Local). To improve the quality of the resulting data, we made conscious decisions as to which, of the many available sources, we should choose and employ in our attempt. The particular decisions are listed along with the data source descriptions, but in general we relied on three factors: (1) the ease by which we could extract the data, i.e., the utility of the site; (2) the quality and amount of data we could obtain; and (3) an a priori understanding of the tasks required to convert a particular type of non-spatial data, i.e., addresses, into spatial data. Regarding the first criterion, we chose sites whose interfaces supported extraction by our agents, which usually indicates a web site that has a database backend, with the data displayed per page as a 'view' of the data based on some search criteria. While a web crawler may be able to identify when this is the case for a particular site it encounters, the ability to automatically determine the context and meaning of the parameters is generally beyond the scope of its capabilities, although advances are being made (Carman and Knoblock 2007).

For the second, we chose sites that were reputable and commercial (to increase the chances of higher quality data because of the vested interest of the site), and large scale (to increase the amount of data we could retrieve). An alternative web crawler-based approach may have also found these sites (and most probably would have because of their large scale and advertising), but in the case of the third criterion,

may not have been capable of realizing the possibility of the link between the non-spatial addresses and their possible conversion into spatial information through the use of other sites and data services. These types of geospatial reasoning processes for analyzing, understanding, and utilizing both the context of a complex geospatial information gathering process (such as gazetteer creation) are currently beyond the bounds of web crawlers, but emerging research is making advances in this direction (Chen and Jing 2004) and the proliferation of geospatial web services will only help move this forward.

Having noted this, our algorithm can be broken down into the following main parts, which we will discuss in detail next:

1. Generate a set of geographic features.
2. Associate a name and type with each feature.
3. Associate a geographic footprint with each feature.

5.1 *Feature generation*

To generate a list of geographic features, we began with the assumption that for an urbanized area, most of the geographic real estate will be covered by addressable data, such as buildings. Using this assumption, if we could generate all possible addresses contained within our urbanized study area, El Segundo, California, we would have a reasonable set of geographic features forming the basis of our gazetteer. Features lacking addresses are omitted using this approach, but the significance of this limitation and possible ways to get past it are taken up towards the end of the paper.

5.1.1 Candidate set. To gather an exhaustive list of addresses one can turn to any number of postal address or street vector data products. Examples of postal address products include the Zip+4 files distributed by the United States Postal Service (USPS) (United States Postal Service 2008b), and other commercial products such as those sold by the Melissa Data Corporation (Melissa Data Corporation 2008). Examples of commonly used vector data products include the US Census Bureau's Topographically Integrated Geographic Encoding and Referencing files (TIGER) (United States Bureau of the Census 2008) and enhanced derivates such as that sold by Tele Atlas (Tele Atlas 2008) or Navteq (NAVTEQ 2008). Each of these formats (postal address and street vector products) includes similar details about the street segments and any could be chosen for a similar project, but one would need to consider the cost, availability, and accuracy required for the application when determining which is appropriate for their needs.

The attributes common to each of these datasets allow one to generate a complete list of all possible addresses for the areas that they cover. The attributes required to perform this task are the street name (for identification) and the valid address ranges on each side of the street. We chose to use the USPS Zip+4 files to generate our addresses because it provided some additional information about what is located at that address if it is known to the USPS (name of occupant, residential or commercial status, etc.). If the information was present, the address would be listed individually in addition to the address range to which it belonged.

To generate an exhaustive list of all possible addresses for our test area, we iterated through each street segment that fell within the geographic region of interest. For each street segment, we then expanded the address ranges on each side of the street, creating an address for each entity within the range. For example if the

address ranges on the two sides of a street were 100–198 and 101–199, we would create the addresses 100, 102, 104 … 198 for one side of the street and addresses 101, 103, 105 … 199 for the other. In the case that there were secondary unit indicators (apartment numbers and suite numbers among others), we would expand them as well. The same scheme was followed to expand non-numerical ranges; for example 123 Main Street Suites A–F was expanded to 123 Main Street Suite A, 123 Main Street Suite B … 123 Main Street Suite F.

Our method of feature generation assumes a consistent address numbering scheme, with one side of the street containing even numbers and the other side being the odd numbers. This assumption holds true for much of the USA and other places which use metric address systems, but would fall short in regions such as Latin America where different address systems or none whatsoever are employed.

5.1.2 Refined set. As Bakshi *et al.* (2004) showed in their work to produce a more accurate geocoder, the addresses produced by expanding the address ranges found in street vector data will include listings for addresses which do not actually exist. This is also the case with the addresses generated by our previously defined method using the USPS ZIP+4 data as the base.

Before adding any address into the gazetteer as a feature, we needed to trim nonexistent addresses from the complete set of candidate addresses produced by the address range expansion. To perform this refinement, one could consult any source which could verify the existence of an address, such as the Geocoded National Address File (G-NAF) database (Paull 2003) for Australia or a rule engine that could reason about the existence of an address as is possible in Denmark where all addresses must follow a strict pattern (Lind 2005).

In Los Angeles County where El Segundo is located, the Los Angeles County Assessor (LACA) website (Office of the Assessor, County of Los Angeles 2008) offers an HTML form which takes as input an address and outputs the Assessor ID Number (AIN) as an HTML link to more details about the property if it exists, or an error if it does not exist. We were able to wrap this form using the AgentBuilder to produce an agent which could verify addresses as either real or nonexistent, and record the AIN for valid addresses. Thus, the input for this agent was an address, and the output was either an AIN or an indication that the address did not exist. We passed each address in the candidate set to this agent, and threw away addresses which the agent determined to be nonexistent. After running all addresses though the verification agent, we had a refined exhaustive set of geographic features (in the form of addresses) for our study area.

Following from our assumption that addressable data constitutes most of the geographic features on the ground in urban areas, the methods outlined in this section would produce an accurate representation of the urban geography for our study area.

5.2 *Name and type association*

So far in the progress of the methodology, the only name or type information associated with each feature has come during the feature generation process. There, if the ZIP+4 data source had any information about the occupant or usage of a particular address indicated by the presence of the address individually, it was recorded along with the feature.

In order to supplement the type information about features which were not identified individually in the USPS Zip+4 files, we again used the LACA site. In response to a successful query for the details page of a particular AIN, the LACA site will return basic information about a property such as its legal description, its worth in the current roll values, and whether the feature was zoned for residential or commercial use. By creating another agent with AgentBuilder to scrape the information on the detail pages about properties, we were able to create a data source which could provide basic information about type of feature, in this case whether it was residential or commercial. We queried this agent using the AINs for each address in the refined set of geographic features previously derived, and thus associated a type of either commercial or residential with each address in our refined set.

We now had the features in our gazetteer separated into two classes, residential and commercial, that still lacked any name information. To obtain more detailed information for each feature (such as a name and a more detailed type), we chose to extract information from online phone books, comprised of large lists of businesses and residences along with their addresses within an area. These are available for most of the world, so using this type of source would allow our approach to be applied to most regions. In addition, online yellow pages, phone books of nonresidential listings, are usually designed for categorical browsing. This means that the authors of the service have designed it in such a way that a user can quickly find what s/he is searching for, represented as hierarchical levels of categories. By traversing through this category structure, one can be assured that the features reached at the terminal branches of the hierarchy (the actual phone number listings) can be associated with the feature types along the path taken to reach them (assuming there are no errors on the site). This implicit typing of features provides a valuable aspect in the creation of a gazetteer, where each geographic feature must be associated with a feature type. Residential phone books, on the other hand, usually provide a single feature type, residences, but do provide rich name information to associate with features.

We identified two such phonebook sites which would be useful for our algorithm; Superpages (Idearc Media Corporation 2008) for commercial addresses because it offered a very nice hierarchical category browsing structure interface, and Switchboard (InfoSpace, Inc. 2008) for residential addresses because it allowed an easy method to search for people's names and addresses, in contrast to most phonebooks which don't offer reverse lookup.

5.2.1 Superpages. To retrieve name and type information about commercial features from the Superpages source, we created two agents with the AgentBuilder software. The first agent we built was responsible for collecting the hierarchy information, while the second gathered the details of features contained in a particular class of the hierarchy. The decision was made to separate these tasks because the agents served independent purposes, one generating the classes used to classify features, and the other gathering name and other attribute information.

The hierarchy retrieval agent can be thought of as a recursive spider. It takes as input a starting page, and extracts the links to sub-hierarchies contained on the page and recursively calls itself for each link. The recursion terminates when there are no other sub-hierarchies on the page, evidenced in the Superpages case by the existence of an HTML form asking the user for a ZIP Code to begin a detailed search within that category. The data extracted and stored about the hierarchy by this agent are:

the name of the category, the unique ID, the depth, and the parent. After running this agent by passing it the webpage corresponding to the top level of the hierarchy, we obtained a complete picture of the ontology used on the Superpages site in addition to every unique ID which could be used to query the site for entities within a specific class of the ontology.

The details retrieval agent was trained to extract the list of features present on the resulting page of a successful class query, as well as navigate through a list of details pages in the case where multiple pages of results were returned. The inputs to this agent were a category ID for a particular class in the Superpages ontology, and a ZIP Code (we used 90245, the ZIP Code of El Segundo). The output of this agent was a list of features separated into the individual non-hierarchical attributes name, phone number, address, website URL, and email address.

Each category ID extracted by the hierarchy agent was passed as input to the details agent to retrieve all entries present in that class of the Superpages hierarchy. The agent would perform its query on the Superpages site and extract the list of features within the input category ID which were within the specified input ZIP Code. The ZIP Code filter was necessary because many 'Advertiser Listing' entries returned from a search for a specific ZIP Code were in fact outside the ZIP Code. Additionally, entries that were extracted without addresses (simply listing phone numbers) were eliminated.

5.2.2 Switchboard. In similar fashion to the Superpages site, we created an agent to query information from the residential phonebook by wrapping the Switchboard site. This site allows one to search for phone numbers and names using an address. The agent that we created for this site took as input the address of a feature and returned as output the name and phone number of the person whom the publicly available phone number was registered to at that address, if one existed. This information does not increase our knowledge about the type of residential features past what we had already known from the LACA site, that they are residential, but it associated a name with each residence which would be valuable for some applications (e.g. emergency response).

After completely running both our Superpages and Switchboard agents, we had effectively materialized a local copy of the Superpages and Switchboard databases for our study area, El Segundo, CA 90245. What this means is that we now had three databases, one of the hierarchical relations defined by Superpages for their entries, one for the entries themselves, and one for the name and phone number information from Switchboard. However, these entries were not yet linked to the geographic features produced in the first step of our algorithm. In the next two subsections we will discuss the operations applied to these data before they could be integrated into our gazetteer.

5.2.3 Normalization. Information extraction tools suffer from the fact that the data they deliver will only be as good as what is available on the site. After data has been extracted using information extraction tools, it is often in a format that we cannot directly work with. Quality assurance and quality control (QA/QC) are sometimes a problem due to errors during extraction, or just plain erroneous data on the site. To rectify this, the data need to be cleaned by recognizing obvious records which are erroneous and either fixing or removing them.

The data we extracted from our Superpages source exemplified this problem in that it did not have any standard format for the addresses. This occurred because

the site allowed businesses to create their own listings and enter their addresses themselves. We chose to transform all of the extracted addresses into the standard USPS addressing format. Once all of the addresses are in a single, well known, uniform format they can easily be transferred into any other desired format. There is a great deal of literature devoted to the address cleaning and standardization process, and the interested reader is directed to Christen and Churches (2005) and the references therein for more information. For our purposes, this process was accomplished by simple token based text processing of the address strings, following the address parsing rules published by the USPS (United States Postal Service 2008a). This technique additionally made use of the ZIP+4 files to determine attributes of addresses such as street suffixes (st., blvd., and so forth), in some cases where this information was missing from the extracted address.

5.2.4 Record linkage. One goal of our reduction process was to create the minimum set of features needed to represent everything that we extracted from our sources, maintaining only a single object for each feature with all of its attributes. Borrowing techniques from the field of record linkage, one can remove duplicates while consolidating the attributes from different instances of the same object (Michalowski *et al.* 2005). While extracting from our Superpages source, if a feature appeared in multiple categories it would be extracted (and end up in the database) multiple times as multiple unique records. A benefit of performing the address normalization process first was that we could easily identify features which were possibly duplicates, based on address. This alone is obviously not strong enough evidence to consider two features the same. However, if we add in the additional constraint that they must have the same name, we can then be reasonably certain that two features which were extracted as distinct are in fact the same. This record linkage may seem a nuisance, but it is actually a side effect of a very desirable characteristic of our Superpages source: the presence of features in more than one category allows us to maintain much richer feature type associations than if they were limited to a single category.

By normalizing the addresses from Superpages, Switchboard, and those created during the feature generation step of the algorithm, we were then able to perform a database join on all three sets with the address as the join key. This process merged the datasets creating a single one with attributes from all sources. The features in the exhaustive set created during feature generation were now associated with the detailed name information extracted from the Superpages detail and Switchboard agents. Additionally, each was linked to one or more specific Superpages classes extracted by the Superpages hierarchy agent in the case of commercial features, or else maintained as residential.

5.3 *Footprint generation*

Once we had acquired a set of unique features from our phonebook sources, each with a name and at least one associated type, we were ready to generate the third and final component of a gazetteer, the footprint. A first approximation for the spatial footprints could be made by simply geocoding their addresses to produce a point footprint for each. There has been a great deal of research into geocoding techniques and technologies in the past few years as people have begun to realize that address data are more useful when they become geospatial data. Several studies have investigated the accuracies associated with different tools and methods of

geocoding (Cayo and Talbot 2003, Fulcomer *et al.* 1998, Ratcliffe 2001, Ward *et al.* 2005, Wieczorek *et al.* 2004, Yang *et al.* 2004, Bonner *et al.* 2003, Whitsel *et al.* 2004, among others), and how these accuracies can have dramatic effects on the outcome of any results based on geocoded data (Krieger *et al.* 2003a, b, McElroy *et al.* 2003, among others).

In brief, a geocoder is any system which can take a piece of non-geospatial information and geographically reference it. The non-geographic information could be an address, a named place, an intersection, or even a textual description of a location. The data sources typically used by geocoding systems include vector data similar to what can be used to generate address ranges (TIGER, Tele Atlas, Navteq, among others), pre-existing address point databases (G-NAF), and even gazetteers. The output of the geocoding process is most commonly, but not strictly limited to, a single geographic point. This output may also contain other information regarding the accuracy of the output, any assumptions that were made during the geocoding process, or any sources which were consulted to achieve the result. The simplest and most commonly employed geocoding methods use linear interpolation along a line segment that has been selected as being the correct segment onto which the address to be geocoded falls. The segment is broken up into equal portions based on the size of the address range, and the centroid of the portion corresponding to the correct address is computed as the geographic output of the geocoder.

This basic process (termed the address range method) has been shown by many researchers to be, at best, challenging and error prone (Bakshi *et al.* 2004, Cayo and Talbot 2003, Lee and McNally 1998, Levine and Kim 1998, McElroy *et al.* 2003, Ratcliffe 2001, among others), and at worst completely infeasible for areas that use non-metric addressing systems such as Japan or Korea (Davis *et al.* 2003). In our previous work (Bakshi *et al.* 2004), we developed an advanced geocoding system capable of utilizing multiple information sources to produce more accurate results. We will briefly outline the improvements to the geocoding process in order to show the similarities between those methods and how we have used similar techniques to generate our refined set of geographic features in the first step of the algorithm. In the previous work, we show that the address range method of geocoding discussed above introduces large amounts of error because it does not take into account the actual number or distribution of lots (parcels) along a street segment during interpolation. The accuracy of the geocoding process can be improved by consulting other sources which are capable of identifying addresses along a segment which actually exist and removing those which do not from the interpolation process, termed the uniform lot method. In the geocoding system developed through our previous work, we utilize the LACA site in a similar fashion as we have done during the feature generation step of the algorithm. In both cases, the LACA site is used to verify the existence of candidate addresses along a street segment. In the geocoder, this information is used to determine a better estimate of the number of parcels on a street segment, and in turn derive closer approximations to actual parcel sizes producing more accurate interpolation results. In the present work, this method is used as a verification service to reduce the candidate set of geographic features to only those which exist.

In addition to utilizing this additional information about the number of parcels on a segment to improve interpolation, our continuing work also demonstrates how a geocoding system can effectively reason about the layout and orientation

of lots to produce more accurate results, termed the actual lot size method. In this method which works for square blocks, the dimensions of the parcels are known but their layout and orientation are not. We make the assumption that parcels on each corner can be oriented towards one of two streets. This leads to 16 possible combinations of parcel layouts and orientations for the parcels which make up the block. For each possibility, the sizes of the parcels on each side of the block are summed and compared to the length of the segment defined in the vector data for that side of the block. The combination which minimizes the error between the calculated length and the actual length of the four segments which make up the block is chosen as the most probable solution. The correct parcel is then identified and its centroid is returned as the geographic output of the geocoder.

To produce a footprint for each of the features currently in our gazetteer, we queried the geocoder using the address associated with each feature. The current geocoder we used did not support sub-parcel level geocoding, so all features at the same address with different secondary unit indicators received the same geocoded coordinates. This process geographically referenced each of the features in our gazetteer, and completed our method for automatically generating the gazetteer database.

After completing this final step of the algorithm, the gazetteer consists of a set of geographic features generated by address range expansion during the first step. Each of these features has an associated name and type that was extracted from our phone book data sources during the second step, and a point footprint (in latitude and longitude) generated by geocoding its associated address in the third step. We recognize that using a simple point representation is not the best footprint that a feature can possibly have, but for the purposes of our experiment, it would suffice in allowing us to determine the feasibility of the novel elements of our overall approach.

6. Results

The aforementioned feature generation method led to the creation of 65,601 candidate addresses, of which there were 56,344 unique addresses (ignoring secondary units) and 9257 which had the same address and differing secondary units. Using the LACA data source to verify addresses, this candidate set was trimmed to produce a refined feature set of 4538 features. The ZIP+4 data sets used in the address generation process provided 144 unique names for 212 unique addresses (ignoring secondary unit distinctions) and 1235 addresses with different secondary units.

The Switchboard source was able to provide name information for 2498 features in the refined feature set. While evaluating the results from the Superpages source, we imposed the restriction that to be a distinct feature required a unique combination of business name and address, because multiple businesses can reside at the same address. From the Superpages source, we gathered 6059 features, of which 4178 were actually within El Segundo. The remaining 1881 were businesses who chose to advertise in El Segundo listings, but were not actually located in El Segundo. The 4178 unique business/address features contained 891 unique addresses. After performing the address normalization previously discussed, the number of unique addresses in the Superpages data was reduced to 689. After the record linkage step was run on the Superpages data, we reduced the number of

unique features it provided information for from 4178 to 1321, essentially removing duplicate information for 2857 features.

6.1 *Feature types*

The type hierarchy for commercial features produced using our process is highly dependent on the source used, in fact mirroring the browsing structure of the Superpages source. We can see that as such, it will not automatically be interchangeable with existing feature type thesaurii (FTT) such as the ADL FTT[1] or the Getty Thesaurus of Geographic Names[2] (TGN). This problem of ontology reconciliation has received significant attention (Hill *et al.* 1999, Doerr 2001, Berman 2003) and we will not go into the details here. Instead we will assume that a solution to this problem exists, and a user of a gazetteer created with our methods could use this solution to perform the translation from one ontology to another.

Each of the 4538 features in the refined set received an initial type indication (either commercial or residential) from the LACA. As previously noted, the Superpages source was able to provide detailed type information for 1321 features. Table 1 summarizes the highest level of the type hierarchy that was constructed from the Superpages source and the number of lower level subcategories and number of features within each category. We can see that even though there are only 1321 unique features, there is a large amount of cross typing, evident from the 7002 features which occur if one were to count every instance of a feature each time it appeared in any category. Another noteworthy fact is that our feature type hierarchy contains 10,239 distinct paths to the leaves.

6.2 *Completeness comparison to ADL and LACBD*

Currently, there is no authority to evaluate the completeness a detailed gazetteer for our test area in terms of features included. However, two gazetteers are available

Table 1. Top level categories of feature types with number of subcategories and features.

Category	Subcategories	Features
Arts & Ent.	309	167
Automotive	560	314
Business & Prof. Serv.	825	821
Clothing & Acces.	247	156
Community & Gov.	427	244
Computers & Elec.	283	413
Construction & Contr.	1097	516
Education	89	127
Food & Dining	460	265
Health & Medicine	774	463
Home & Garden	1092	607
Industry & Agric.	1476	416
Legal & Financial	224	485
Media & Comm.	321	248
Personal Care & Serv.	104	310
Real Estate	166	333
Shopping	668	439
Sports & Rec.	521	149
Travel & Trans.	596	529
Totals	10,239	7002

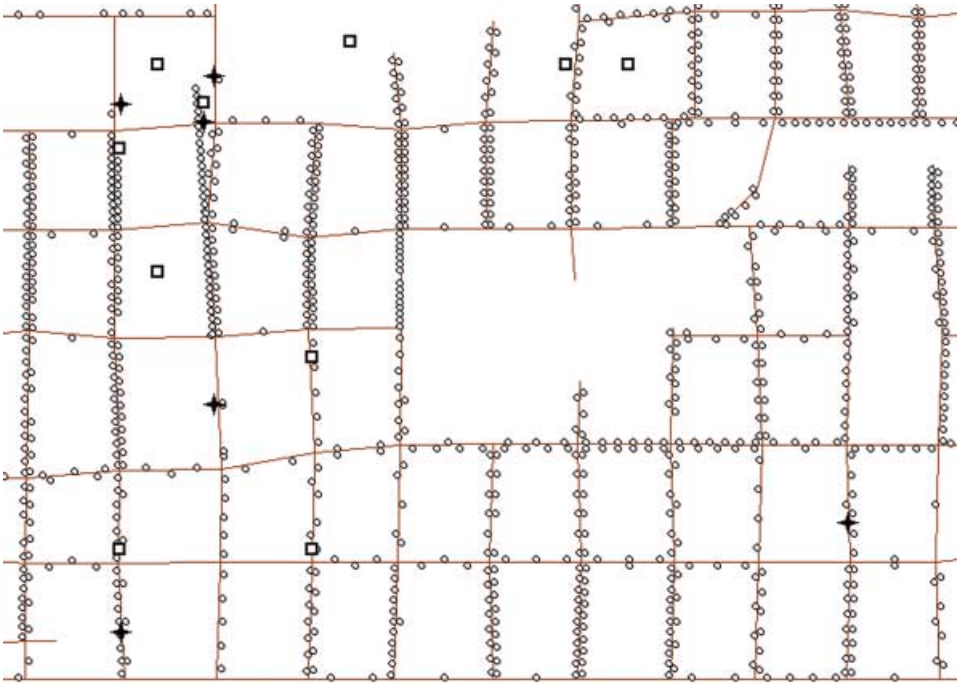

Figure 1. Features contained in the ADL (squares), LACBD (stars), and AGG (circles) for a sub-region of El Segundo.

which cover our test area that we can compare ours against: the ADL[3],considered to be the most complete gazetteer for the USA, and the Los Angeles Comprehensive Bibliographic Database[4] (LACBD), considered to be the most detailed gazetteer for Los Angeles County. Figure 1 displays the features present within a sub-region of El Segundo bounded by Concord St to the west, Palm Ave to the north, Penn St to the east, and Grand Ave to the south for the ADL, LABCD, and the automatically generated gazetteer (AGG) formed by combining the extracted data from our three online data sources (the Superpages, Switchboard, and Assessor sites). The subregion shown contains the highest density of features for both the ADL and LACBD within El Segundo.

In order to perform an evaluation of the completeness of the AGG, we combined the features contained in both the ADL and LACBD gazetteers to create a super-set of features. This was treated as the best knowledge we had about our area, and used as a basis to determine the recall of each gazetteer. This is far from an ideal comparison, but it allowed us see how well our algorithm performs based on the gazetteers that are currently available. Table 2 lists this complete superset of features with their types, indicating which features are present in which gazetteer, and showing the recall ratio and percentage for each gazetteer.

The ADL is in fact more detailed (in terms of the number of features present) than the LACBD, but neither lists many features for the test area. Of the combined superset, the ADL provides 83% of the entries, while the LACBD provides 38%. The AGG covers 63% of the points in the combined ADL and LACBD gazetteers. However, it should be noted that the AGG is, by design, a gazetteer of only man-made structures as it is based on features generated from street addresses and

Table 2. Comparison of features included in each gazetteer using the following types: B, building; E, educational facility; L, library; PO, post office; R, religious facility; T, theater; H, heliport; PK, park; and S, sports facility. X indicates existence, – indicates non-existence.

Type	Name	ADL	LACBD	AGG
	Buildings and Subtypes of Building			
B	El Segundo CityHall	X	X	X
E	Center Street Elementary	X	X	X
E	St Johns Lutheran	X	–	X
E	St Anthony Elementary	–	X	X
E	Richmond Street Elementary	–	X	X
E	Webster University	–	X	X
E	Arena High School	X	–	–
E	El Segundo High School	X	X	X
E	El Segundo Middle School	X	X	X
L	El Segundo Public Library	X	X	X
PO	El Segundo Post Office	X	–	X
R	Pacific Baptist Church	X	–	X
R	St Andrew Russian-Byzantine Catholic Church	X	–	X
R	Temple Rodeph Shalom	X	–	X
R	El Segundo Christian Church	X	–	X
R	Foursquare Church Of El Segundo	X	–	X
T	Old Town Music Hall	–	X	–
	Non-Buildings			
H	Chevron	X	–	
H	Airport Towers #1	X	–	–
PK	Holly Valley Park	X	–	–
PK	Candy Cane Park	X	–	–
PK	El Segundo Park	–	X	–
PK	Kansas Park	X	–	–
S	El Segundo Golf Course	X	–	X
Ratio of total (superset)	20/24	9/24	15/24	
Recall % (superset)	83%	38%	63%	
Ratio of total (buildings)	13/17	9/17	15/17	
Recall % (buildings)	76%	53%	88%	

phone numbers. Taking this into account, the AGG did better (88%) than both the ADL (76%) and LACBD (53%) sources for the combined superset comprised of only features of type building and its subclasses (educational facility, library, post office, religious facility, and theater).

An investigation into features not present in the AGG revealed an incorrect entry in the ADL within our study area. In this case, erroneous geocoding methods identified the New Mount Calvary Missionary Baptist Church, located at 402 East El Segundo Blvd, Los Angeles, CA 90061 as falsely being located at 402 East El Segundo Blvd, El Segundo, CA 90245. Here, we can speculate with reasonable certainty that the geocoding process used selected the wrong street segment to use as a basis for the interpolation of the point, using the 400 block of El Segundo Blvd in El Segundo, 90245 instead of the correct 400 block of El Segundo Blvd in Los Angeles, 90061. This was manually confirmed by geocoding the address using both segments as a basis for interpolation [following the methods detailed in Bakshi *et al.* (2004)], and a comparison of the resulting positions identified the nearest street segment to the position of the feature listed in the ADL to be the 400 block of El

Segundo Blvd, in El Segundo, 90245. From this we can be reasonably certain that this (incorrect) segment in El Segundo was selected as the basis for geocoding, based on: (1) the similarity of the street names of the two blocks; (2) the similarity of the address ranges of the two blocks (400–499 for both segments); and (3) the proximity of the coordinates of the ADL feature to that of the (incorrect) street segment. Our AGG was able to identify the correct location of this feature (outside of El Segundo), and it is not present in our results. Hence, when comparing the AGG and ADL to the combined superset, the AGG gazetteer is more precise than the ADL in that it contained no incorrect features, as opposed to the incorrect feature found in the ADL.

Similarly, the LACBD contains incorrect/outdated information for the feature Temple Rodeph Shalom. The LACBD maintains the footprint of this feature as a point, while the feature in the AGG does not have a footprint at all. This is because it was extracted from Superpages with a USPS Post Office Box which is not geocodable, and therefore has no footprint associated with it. A phone call to the organization verified that their address is a Post Office Box, because they no longer have a physical location of their own, instead using a meeting room at a local hotel. The AGG was correct in not storing a footprint for this feature. We should note it could be argued that the Temple Rodeph Shalom is not a feature at all because it does not have all three components of a gazetteer feature: the name, type, and footprint. The AGG outperforming a gazetteer created with local level knowledge, the LACBD, in terms of false positives is an unexpected result.

6.3 *Spatial accuracy comparison to ADL and LACBD*

The preceding section focuses on evaluating the completeness of the AGG in terms of the features it contains for an area. Another equally important aspect of any gazetteer is the spatial accuracy of the footprints it maintains for its features. To measure this, the footprints from the ADL, LACBD, and AGG for the features in the combined superset, limited to only those of type building and its subtypes, were displayed on top of high resolution satellite imagery, and the distance to the actual centre of the feature in the imagery was computed. The actual coordinates of each of the features and the spatial accuracies in terms of distance to the actual features of the three gazetteers are listed in table 3 (feature names from table 2 are abbreviated and feature type information is omitted to save space). In the case where a feature is comprised of multiple buildings, the centre of the feature is assumed to be the centre of the parcel on which they reside. The feature Temple Rodeph Shalom is excluded from this evaluation because it has a Post Office Box address, instead of an actual physical street address that can be geocoded. The results show that the AGG is the most spatially precise gazetteer, with gazetteer features located 49.62 m on average from the actual centre of the feature as seen on the image. The high spatial accuracy of the AGG is due to the high spatial accuracy of the underlying geocoder used to produce the footprints.

6.4 *Ground truth evaluation*

In addition to being able to measure the completeness and spatial accuracy of our AGG by comparing it to existing gazetteers, a more useful measure would be how well the AGG represents what is actually on the ground. Therefore, we chose 43 street segments representing various geographic feature types (mainly industrial (I),

Table 3. Spatial accuracy of each gazetteer in terms of distance (m) to actual feature location (lat, lon).

Name	Actual	ADL	LACBD	AGG
City Hall	33.9199, −118.4153	70.79	48.28	41.36
Center St	33.9250, −118.4036	34.54	91.91	94.88
St John's	33.9285, −118.3983	65.49	–	41.50
St Anthony	33.9185, −118.4083	–	41.18	42.93
Richmond St	33.9243, −118.4177	–	61.64	64.42
Webster	33.9147, −118.3760	–	668.51	–
Arena	33.9247, −118.4123	75.08	–	–
High School	33.9243, −118.4148	56.92	83.59	80.26
Middle School	33.9206, −118.4036	1025.74	101.54	94.59
Library	33.9238, −118.4182	157.09	157.87	184.92
Post Office	33.9180, −118.4155	63.85	–	21.75
Pacific Baptist	33.9286, −118.4163	53.09	–	24.63
St Andrew	33.9234, −118.4180	61.73	–	23.81
Christian Church	33.9180, −118.4179	59.49	–	15.43
Foursquare	33.9217, −118.4174	58.72	–	18.21
Old Town	33.9175, −118.4168	–	57.76	45.26
Average distance		111.41	82.02	49.62

commercial (C), residential (R), and mixes of these (I/C, C/R)) and compared what the gazetteer sources claimed was there with what was actually there by walking the blocks and visually identifying buildings. Figure 2 shows an aerial image of El Segundo, with all road segments in white and the road segments included in the survey overlaid in black.

Figure 2. Aerial of image El Segundo with all road segments in white and road segments included in ground survey overlaid in black.

To measure the performance of our methods, we extend the traditional notions of precision and recall of van Rijsberg (1979) to define the following terms:

- *Location Precision*: How many of the features we extracted actually exist?
- *Location Recall*: How many of the total number of features in existence did we extract?
- *Name/Type Precision:* How many of the names and types that we extracted for features were correct?
- *Name/Type Recall:* How many did we get correct of the possible names and types that actually exist?

In the following subsections, we will use these measures to evaluate and discuss the relative strengths and weaknesses of each of the sources used to generate our gazetteer. Table 4 lists the location precision and recall for the gazetteer produced from the data obtained from each source and that of all three sources combined.

6.4.1 Superpages source. The gazetteer produced using the Superpages source has an overall location precision of 72% and recall of 58%. However, this source often overestimates the existence of distinct geographic features (buildings) in two situations. The first is dense commercial areas such as downtown street segments where multiple storefronts share the same building. Here, the Superpages will produce several distinct geographic features (buildings) when they are, in fact, all part of the same building. The second area of overestimation occurs in commercial plazas made up of a single structure around a parking lot. This again is a single building, with multiple addresses, leading to the appearance of multiple buildings in the gazetteer. These two factors lead to low precision in terms of geographic features (buildings) in dense commercial areas.

These problems raise a more general question for gazetteers, particularly evident in those created using our approach, which we will not attempt to answer here: should the granularity and/or type of feature be determined by physical structure or its occupants and/or uses? One could argue that the physical structure is an appropriate focus, in which case all buildings would be assigned a single type (schools, homes, police stations, and so on). On the other hand, if the occupants of buildings constitute the appropriate level of granularity, the number of types in Table 1 suggests that we would need a feature type hierarchy with over 10,000 different entries. Hill and Zheng (1999) maintain that all named geographic places are proper features to represent in a gazetteer, but at what level of detail a geographic place should be defined is still an open issue. What actually occurs in practice today is a mixture of the two approaches, referring to a police station as a different type than an educational facility even though they may both be public

Table 4. Location precision/recall percentages of each data source by region type.

Region type	Superpages	Switchboard	Assessor	Combined
I	0.90/0.39	0.10/0.08	0.86/0.71	0.94/0.08
I/C	0.88/0.78	0.00/0.00	0.94/0.84	0.92/1.00
C	0.75/0.99	0.67/0.17	0.84/0.72	0.64/1.00
C/R	0.94/0.63	0.78/0.45	0.86/0.89	0.79/0.99
R	0.00/0.00	1.00/0.44	1.00/0.91	0.99/0.91
Overall	0.72/0.58	0.49/0.22	0.89/0.81	0.85/0.94

Table 5. Name/type precision and recall of Superpages data by region type.

Type	Precision	Recall
I	1.00	0.68
I/C	1.00	0.83
C	0.95	0.88
C/R	0.92	0.87
R	N/A	N/A
Overall	0.96	0.81

buildings, while not distinguishing between the different types of commercial sites based on the occupants present.

Table 5 lists the name/type precision and recall for the Superpages source. The fact that there is nearly 100% name/type precision means that the yellow pages data are almost always correct in terms of the names and types of businesses located at addresses. One potential reason for the high precision of this source is that the businesses paid to list themselves. Hence, they have a vested interest in verifying that information returned about their business is accurate. As previously stated, the Superpages source contains only commercial data, and as such the name/type precision and recall are not applicable to strictly residential areas because there are very few commercial sites located there.

One final weakness of the Superpages source is that only commercial businesses typically looking for walk-in customers are represented. For instance, warehouses that are not the main office for a business are typically missing. This leads to the low recall observed in industrial areas.

6.4.2 Switchboard source. As a residential phonebook, it is appropriate that this source provides very precise data for residential areas. However the recall for the residential areas was lower than expected. Possible reasons for this could include unlisted phone numbers and the fact that more and more people rely solely on cell phones instead of traditional land lines. The recall and the precision were lower in other region types as expected. This source, for example, did not do well in commercial or industrial areas because there are few residences present in these areas.

As with the Superpages source, this source also raises a more general question for gazetteers: should residence information be included in a gazetteer, and if so what are the privacy concerns that may go along with it? One could simply include the feature as a residence type and not the name of the owner of the telephone, but for certain applications such as police and fire departments, this personal information may be required so it might need to be included. The data used for this paper has been derived from public sources, so perhaps privacy is not so much an issue in this particular case.

6.4.3 Assessor source. This source was the most precise in terms of the actual geographic features that it produced. It rarely overestimated the number of features, more often underestimating them in the case where multiple entities are located on the same parcel of taxable property. The extremely high accuracy available from this source can be traced to the fact that the underlying data was created by the local government to aid the Assessor's Office in real estate tax collection. For this reason, there is a strong desire to levy the correct taxes against each landowner. One

drawback of this source is that it did not provide very detailed type information about the features that it produced.

6.4.4 Combined sources. By combining the data from our three sources, we were able to exploit the particular advantages of each data source, and the resulting gazetteer had very good recall and fairly high precision across all types of regions. This combined gazetteer contained nearly all buildings which exist on the ground in our study area. Additionally, the precision is also quite good in noncommercial areas, providing very few false positives.

7. Discussion

As previously stated, the resulting gazetteer created using our methods is highly representative of the sources that are used. In our case, we used sources which contained information about features that had addresses, and as such generated a highly detailed gazetteer focused on manmade structures. This type of gazetteer is useful for urban areas where most of the real estate on the ground is in fact covered by buildings, but would fall short for rural or undeveloped areas. This set of current sources is not able to capture and represent natural features such as mountains, rivers, deserts, or oceans. Similarly, our current set of sources provides only the simplest of geographic footprints, latitude/longitude points, and a very limited amount of temporal information can be derived from the source, namely that the feature existed at the time of extraction, with no notion of temporal extent.

The feature types in our gazetteer are not linked automatically to the feature type lists commonly used in gazetteers. Our method does not map the extracted types to one or more existing feature type thesaurii (FTT) so that the types associated with features are those which it appeared under when it was extracted. Additionally, the feature types associated with each feature are typically very specific, more than what would traditionally fit into a FTT, and this may be a significant shortcoming for an application wishing to use our gazetteer directly and/or enable the easy and automatic expansion and integration of existing FTT. The development of FTT and reconciliation of different sources needs more research.

The most serious drawback of our approach is the same limitation common to any information extraction algorithm, given that it is completely reliant on the sources used. No matter what source is used, if there is an error in that source, the error will appear in the resulting data (unless other sources are used for validation). Similarly, if the source is not complete, the resulting data will not be complete.

8. Related work

Some groups have utilized automated approaches to combining two or more gazetteers into a single unified gazetteer (Hill *et al.* 1999, Axelrod 2003). This consolidation process raises several issues concerned with ontology reconciliation and feature resolution that are the focus of a great deal of research in both the geographic and non-geographic domains (Sintichakis and Constantopoulos 1997, Hill *et al.* 1999, Doerr 2001, Berman 2003, Agarwal 2004, Leidner 2004b, Kokla and Kavouras 2005, Brodaric and Gahegan 2006).

Most of the research on creating gazetteer feature databases automatically from scratch has been conducted by the natural language processing (NLP) community out of their need for detailed gazetteers to identify entities in text using named entity

recognition (NER) systems. Many researchers now use the Internet as a source of named entity information (Ferres *et al.* 2004a, b, Maynard *et al.* 2004, Pekar and Evans 2005). However, none of these NLP approaches adhere to the strict definition of a gazetteer as defined by Hill (2000), where a feature must have an associated name, type and footprint. Instead, these works each focus on a single aspect of the gazetteer, either building a complete thesaurus of entity names, or refining either the type or footprint information. A gazetteer produced using these methods would work well for the NER community where the task is simply to identify names of locations, but for other fields where the full definition of a gazetteer is required, it would not suffice.

The work most relevant to ours is that of Uryupina (2002, 2003) where text classifiers and patterns are learned for a limited set of feature types (for example city, region, country, island, river, and mountain). These patterns are used to search the Internet, and based on the count of pages that come back, feature types are identified. This approach is similar to ours in that they are exploiting the Internet to help with the creation of a gazetteer, but they are not explicitly extracting their features from the Internet, instead using it as a source to determine if something is a geographic feature, and if so, of what type. This approach is limited in terms of the number of feature types for which it will work, and because the generated features lack footprints.

It has long been recognized that the Internet is a rich source of geographic information. Many researchers have proposed methods to determine the geographic scope of web pages to enable geographic search engines, sometimes known as geospatial Internet browsing, geographic information retrieval (GIR), or local search (Buyukkokten *et al.* 1999, Clough 2005, Ding *et al.* 2000, Himmelstein 2005, Martins *et al.* 2005b, McCurley 2001, Riekert 2002, Zong *et al.* 2005), and other work has attempted to extend the geographic scope from its traditional point form to more detailed areas such as polygons (Schlieder *et al.* 2001, Schockaert *et al.* 2005), and enabling approximate spatial footprint representations (Jones *et al.* 2001, Wilson *et al.* 2004). When performing place name detection these methods typically rely on NLP and NER techniques which utilize gazetteers to identify geographic entities for them to index and thus create efficient searching algorithms. Without the gazetteers the methods described in this paper are capable of creating, GIR is severely limited in its usefulness because it is the gazetteer which helps to power GIR by identifying and providing the spatial grounding for named entities in the textual documents. Other approaches locate and extract postal addresses from text to georeference the document using a geocoder (Morimoto *et al.* 2003, Sagara *et al.* 2001, Wang *et al.* 2005). This is similar to our approach in that address data is exploited to generate geographic scope, but we are unaware of any other work which attempts to generate a gazetteer from this information.

Likewise, previous researchers have recognized that phonebook information can be exploited generate geographic datasets, but to our knowledge no work has been done to transform them into a classical gazetteer in any sense according to Hill's definition as a 'dictionary or list of geographic names, together with their geographic locations, their feature types, and other descriptive information' (Hill 2006: 228), with specific rules and relationships between and within the data structures and geographic objects. For complete details on the standard description of the classical modern gazetteer, one should consult chapter five of Hill (2006). Lee and McNally (1998) introduce the idea of geocoding the addresses from a phone

book for activity-based travel forecasting which requires the use of the business type information. Our research takes a similar approach in that we also use phone book information to derive type information of geographic features, but in their case they argue that one should obtain local copies of specialized phone book databases from vendors, an assumption which we do not make. Our work focuses on ways in which one can use a variety of readily available heterogeneous data sources to gather and integrate the information needed during the automated gazetteer creation process.

Our extraction and integration approach to automatically generating a gazetteer uses many of the same extraction tools, techniques and data sources as that of Chen *et al.* (2003) and Bakshi *et al.* (2004). However, neither of these works focused on generating a gazetteer from the data they were extracting. Instead, the geographic data they extracted were used directly in their applications without further processing to turn it into a reusable gazetteer in the classic sense of Hill (2006, Chapter 5). In Bakshi *et al.* (2004), the work focused on using similar data sources for extracting information about address ranges along street segments for the purposes of improving and eliminating the assumptions used in the geocoding process, resulting in more accurate geocodes for addresses. In Chen *et al.* (2003), the work focused on utilizing similar sources for the purposes of aligning (conflating) vector data sources with imagery to improve the location of geocodes, thus improving the location of possible sites for buildings. The work presented here can be considered an extension to this existing body in that the gazetteer generated by our methods would be useful as a component in each of their systems, relieving them of the burden of extracting the data themselves and improving the attributes they can harvest from the data. Further, our work takes the next logical step by creating a more complete and accurate picture of the landscape along all three axes of the gazetteer, not just the geospatial footprints that were the focus of our previous work.

9. Conclusions and future work

In this paper we have demonstrated the need for very detailed gazetteers in diverse research arenas, and have outlined an algorithm which can be used to rapidly and automatically create a detailed gazetteer from Internet sources. Our algorithm has the strength that it is general and it could be employed to create a gazetteer for any area for which similar data sources are available. The resulting gazetteer produced by these methods will be very dependent on the sources used in terms of feature types represented and completeness. The new gazetteer described here is richer than existing gazetteers in the test area with regard to common or overlapping feature types, as we have shown through the comparison with two existing gazetteers. Finally and perhaps most importantly, the ground survey measured precision and recall, and shows that our methods (as well as other methods based on the same information extraction approaches) can and should be used to create detailed regional gazetteers with high levels of precision and completeness rapidly and automatically.

We hope to extend our algorithm to include sources which will allow us to automatically include other feature and footprint types. Promising directions include extracting features from online maps and incorporating existing datasets which are available online such as hydrological and census data. In addition, we are investigating methods to extract the actual parcel (polygon) from the images on the LACA site as well as high resolution aerial photos (Ming *et al.* 2005). Successfully doing this would provide very detailed polygon footprints for the parcels to which

an address belongs. We could then overlay these polygon parcel boundaries on a satellite image and use image processing techniques to identify the actual buildings per parcel. With a physical building count per parcel we could possibly reduce the overestimation of the Superpages source by realizing that a set of addresses that do not exist in the Assessor data are actually addresses underneath the same roof, and exploit the excellent typing information of the Superpages along with the good building precision of the Assessor site.

In order to facilitate our gazetteer being easily adopted into existing applications, we need to enable the automatic conversion of our extracted feature types into more standard and traditionally used feature types. We therefore intend to try to incorporate some of the ontology integration work published by other researchers into our gazetteer creation process, enabling the automatic association of more general feature types along with our highly detailed versions.

Acknowledgments

This research is based upon work supported in part by the National Science Foundation under Award No. IIS-0324955, in part by the Science, Mathematics, and Research for Transformation (SMART) Defense Scholarship for Service Program, and in part by the University of Southern California Libraries. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

Notes

1. http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302.
2. http://www.getty.edu/research/conducting research/vocabularies/tgn.
3. http://www.alexandria.ucsb.edu/clients/gazetteer.
4. http://www.usc.edu/isd/archives/arc/lacbd/geographic/.

References

AGARWAL, P., 2004, Contested nature of place: Knowledge mapping for resolving ontological distinctions between geographical concepts. In *Geographic Information Science*, M.J. Egenhofer, C. Freksa and H.J. Miller (Eds), Vol. 3234, pp. 1–21 (Berlin: Springer-Verlag).

AGOURIS, P., BEARD, K., MOUNTRAKIS, G. and STEFANIDIS, A., 2000, Capturing and modeling geographic object change: A spatiotemporal gazetteer framework. *Photogrammetric Engineering and Remote Sensing*, **66**(10), pp. 1224–1250.

ALEXANDRIA DIGITAL LIBRARY, 2006, Alexandria digital library content standard. Available online at: http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2guide.htm (accessed 29 January 2008).

ALEXANDRIA DIGITAL LIBRARY, 2008, Alexandria digital library gazetteer. Available online at: http://www.alexandria.ucsb.edu/clients/gazetteer (accessed 29 January 2008).

AMITAY, E., HAR'EL, N., SIVAN, R. and SOFFER, A., 2004, Web-A-Where: Geotagging web content. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR '04)*, M. Sanderson, K. Järvelin, J. Allan and P. Bruza (Eds), pp. 273–280 (New York: Association for Computing Machinery).

AUERBACH, R., 2006, Consideration of assessor's electronic parcel map. *GIS data Distribution Policy Letter*. Available online at: http://www.opendataconsortium.org/documents/LAC Change Parcel Policy.pdf (accessed 29 January 2008).

AXELROD, A., 2003, On building a high performance gazetteer database. In *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*, A. Kornai and B. Sundheim (Eds), pp. 63–68 (Boston, MA: Association for Computational Linguistics).

BAKSHI, R., KNOBLOCK, C.A. and THAKKAR, S., 2004, Exploiting online sources to accurately geocode addresses. In *Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS '04)*, D. Pfoser, I.F. Cruz and M. Ronthaler (Eds), pp. 194–203 (Washington, DC: Association for Computing Machinery).

BARCLAY, T., SLUTZ, D.R. and GRAY, J., 2000, TerraServer: A spatial data warehouse. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, W. Chen, J.F. Naughton and P.A. Bernstein (Eds), pp. 307–318 (New York: Association for Computing Machinery).

BEACH, J., MINTON, S. and RZEPKA, W., 2004, A Software agent infrastructure for timely information delivery. In *Knowledge Sharing and Collaborative Engineering*, M. Boumedine and S. Ranka (Eds) (St. Thomas, US Virgin Islands: ACTA). Available online at: http://www.actapress.com/PaperInfo.aspx?PaperID=17261 (accessed 29 January 2008).

BEAMAN, R., WIECZOREK, J. and BLUM, S., 2004, Determining space from place for natural history collections: In a distributed digital library environment. *D-Lib Magazine*, **10**(5). Available online at: www.dlib.org/dlib/may04/beaman/05beaman.html (accessed 29 January 2008).

BERMAN, M.L., 2003, Semantic interoperability and cultural specificity: Examples from Chinese, Japanese, Mongolian and Uighur. Unpublished, presented at *the 28th Annual Meeting of the Social Science Association*, Baltimore, MD, USA, 13–16 November 2003.

BERMAN, M.L., 2004, Key issues in compiling a digital gazetteer for China's historical religious sites. Unpublished, presented at *ECAI Congress of Cultural Atlases: The Human Record*, Berkeley, CA, USA, 7–10 May 2004.

BONNER, M.R., HAN, D., NIE, J., ROGERSON, P., VENA, J.E. and FREUDENHEIM, J.L., 2003, Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*, **14**(4), pp. 408–411.

BRODARIC, B. and GAHEGAN, M., 2006, Representing geoscientific knowledge in cyberinfrastructure: Challenges, approaches and implementations. In *GeoInformatics, Data to Knowledge, Geological Society of America Special Paper 397*, A.K. Sinha, (Ed.), pp. 1–20 (Boulder, CO: Geological Society of America).

BUCKLAND, M. and LANCASTER, L., 2004, Combining place, time, and topic. *D-Lib Magazine*, **10**(5). Available online at: www.dlib.org/dlib/may04/buckland/05buckland.html (accessed 29 January 2008).

BUYUKKOKTEN, O., CHO, J., GRACA-MOLINA, H., GRAVANO, L. and SHIVAKUMAR, N., 1999, Exploiting geographical location information of web pages. In *Proceedings of The Second International Workshop on the World Wide Web and Databases (WebDB'99)*, S. Cluet and T. Milo (Eds), pp. 91–96 (New York: Association for Computing Machinery).

CARMAN, M.J. and KNOBLOCK, C.A., 2007, Learning semantic descriptions of web information sources. In *IJCAI 07*, M.M. Veloso, (Ed.), pp. 2695–2700 (Hyderabad: IJCAI).

CAYO, M.R. and TALBOT, T.O., 2003, Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*, **2**(10), pp. 10–21.

CHAVEZ, R.F., 2000, *Generating and Reintegrating Geospatial Data*, pp. 250–251 (New York: Association for Computing Machinery).

CHEN, C.C., KNOBLOCK, C.A., SHAHABI, C. and THAKKAR, S., 2003, Building finder: A system to automatically annotate buildings in satellite imagery. In *Proceedings of the*

International Workshop on Next Generation Geospatial Information (NG2I '03), P. Agouris, (Ed.) (Cambridge, MA: Harvard University).

CHEN, C.C., KNOBLOCK, C.A., SHAHABI, C., THAKKAR, S. and CHIANG, Y.Y., 2004, Automatically and accurately conflating orthoimagery and street maps. In *Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS '04)*, D. Pfoser, I. FCruz and M. Ronthaler (Eds), pp. 47–56 (Washington, DC: Association for Computing Machinery).

CHEN, L. and JING, N., 2004, *A Dynamic Data Structure FOR Geospatial Web Services Integration*, pp. 800 (Washington, DC: IEEE Computer Society).

CHO, G.C.H., 2007, Geographic information, personal privacy, and the law. In *Handbook of Geographic Information Science*, J.P. Wilson and A.S. Fotheringham (Eds), pp. 519–539 (Oxford: Blackwell).

CHRISTEN, P. and CHURCHES, T., 2005, A probabilistic deduplication, record linkage and geocoding system. In *Proceedings of the ARC Health Data Mining workshop* (Canberra, AU: The Australian National University), pp. 109–116.

CLOUGH, P., 2005, Extracting metadata for spatially-aware information retrieval on the Internet. In *Proceedings of the 2005 ACM Workshop of Geographic Information Retrieval (GIR'05)*, C. Jones and R. Purves (Eds), pp. 17–24 (New York: Association for Computing Machinery).

DAVIS, C.A., FONSECA, F.T. and DE VASCONCELOS BORGES, K.A., 2003, A flexible addressing system for approximate geocoding. In *Proceedings of V Brazilian Symposium on GeoInformatics* (Sao Paulo, Brazil, November).

DAWES, S., COOK, M. and HELBIG, N., 2006, *Challenges of Treating Information as a Public Resource: The Case of Parcel Data*, Vol. 4, pp. 1–10 (Piscataway, NJ: IEEE).

DING, J., GRAVANO, L. and SHIVAKUMAR, N., 2000, Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, A.E. Abbadi, M.L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter and K.Y. Whang (Eds), pp. 545–556 (San Francisco, CA: Morgan Kaufmann).

DOERR, M., 2001, Semantic problems of thesaurus mapping. *Journal of Digital Information*, **1**(8). Available online at: jodi.tamu.edu/Articles/v01/i08/Doerr/ (accessed 29 January 2008).

DOERR, M. and PAPAGELIS, M., 2007, A method for estimating the precision of placename matching. *IEEE Transactions on Knowledge and Data Engineering*, **19**(8), pp. 1089–1101.

ELECTRONIC CULTURAL ATLAS INITIATIVE (ECAI), 2008, Available online at: http://www.ecai.org (accessed 29 January 2008).

FERRES, D., MASSOT, M., PADR, M., RODRGUEZ, H. and TURMO, J., 2004a, Automatic building gazetteers of co-referring named entities. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC' 04)*, M.T. Lino, M.F. Xavier, F. Ferreira and R.S.R. Costa (Eds) (Paris: European Language Resources Association).

FERRES, D., MASSOT, M., PADR, M., RODRGUEZ, H. and TURMO, J., 2004b, Automatic classification of geographical named entities. In *Proceedings 4th International Conference on Language Resources and Evaluation (LREC '04)*, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa and R. Silva (Eds) (Paris: European Language Resources Association).

FETCH TECHNOLOGIES, INC., 2008, Fetch AgentBuilder. Available online at: http://www.fetch.com (accessed 29 January 2008).

FONDA-BONARDI, P., 1994, *House numbering systems in Los Angeles*, pp. 322–331, (Phoenix, AZ: Urban and Regional Information Systems Association).

FULCOMER, M.C., BASTARDI, M.M., RAZA, H., DUFFY, M., DUFFICY, E. and SASS, M.M., 1998, Assessing the accuracy of geocoding using address data from birth certificates: New Jersey, 1989 to 1996. In *Proceedings of the third National Geographic Information*

Systems in Public Health Conference, San Diego, CA, USA, 18–20 August, pp. 547–560.

GOODCHILD, M., 1999, The future of the gazetteer. Unpublished, presented at the *Digital Gazetteer Information Exchange Workshop*, Washington, DC, USA, 13–14 October.

GOOGLE, INC., 2008, Google maps. Available online at: http://maps.google.com (accessed 29 January 2008).

HEAL THE BAY, 2008, Beach report card – Los Angeles county. Available online at: http://www.healthebay.org/brc/grademap.asp?map=3 (accessed 29 January 2008).

HENSON, K.M. and GOULIAS, K.G., 2006, Preliminary assessment of activity and modeling for homeland security applications. In *Transportation Research Record: Journal of the Transportation Research Board*, pp. 23–30 (Washington, DC: Transportation Reserach Board).

HILL, L.L., 2000, Core elements of digital gazetteers: Placenames, categories, and footprints. In *Proceedings of Research and Advanced Technology for Digital Libraries, 4th European Conference (ECDL '00)*, J.L. Borbinha and T. Baker (Eds), Vol. 1923, pp. 280–290 (London: Springer).

HILL, L.L., FREW, J. and ZHENG, Q., 1999, Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, **5**(1). Available online at: www.dlib.org/dlib/january99/hill/01hill.html (accessed 29 January 2008).

HILL, L.L. and ZHENG, Q., 1999, Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, pp. 57–69 (Medford, NJ: Information Today).

HILL, L.L., 2006, *Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing)* (Cambridge, MA: MIT Press).

HIMMELSTEIN, M., 2005, Local search: The Internet is the yellow pages. *Computer*, **38**(2), pp. 26–34.

IDEARC MEDIA CORPORATION, 2008, SuperPages: Yellow pages & white Pages. Available online at: http://www.superpages.com (accessed 29 January 2008).

INFOSPACE, INC., 2008, Yellow pages, white pages, maps, and more -switchboard. Available online at: http://www.switchboard.com (accessed 29 January 2008).

JANEE, G., FREW, J. and HILL, L.L., 2004, Issues in georeferenced digital libraries. *D-Lib Magazine*, **10**(5). Available online at: www.dlib.org/dlib/may04/janee/05janee.html (accessed 29 January 2008).

JONES, C.B., ALANI, H. and TUDHOPE, D., 2001, Geographical information retrieval with ontologies of place. In *Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science*, D.R. Montello, (Ed.), Vol. 2205, pp. 322–335 (London: Springer-Verlag).

KNOBLOCK, C.A., LERMAN, K., MINTON, S. and MUSLEA, I., 2000, Accurately and reliably extracting data from the web: A machine learning approach. *IEEE Data Engineering Bulletin*, **23**(4), pp. 33–41.

KOKLA, M. and KAVOURAS, M., 2005, In *Semantic Information in Geo-Ontologies: Extraction, Comparison, and Reconciliation*, Vol. 3534, pp. 125–142 (Berlin: Springer).

KRIEGER, N., CHEN, J.T., WATERMAN, P.D., SOOBADER, M.J., SUBRAMANIAN, S.V. and CARSON, R., 2003a, Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? *American Journal of Epidemiology*, **156**(5), pp. 471–482.

KRIEGER, N., WATERMAN, P.D., CHEN, J.T., SOOBADER, M.J. and SUBRAMANIAN, S.V., 2003b, Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: Geocoding and choice of area-based socioeconomic measures. *Public Health Reports*, **118**(3), pp. 240–260.

LAENDER, A., RIBEIRO-NETO, B., SILVA, A. and TEIXEIRA, J., 2002, A brief survey of web data extraction tools. *SIGMOD Record*, **31**(2), pp. 84–93.

LAWSON, C., 2005, Data fusion. *Transportation Research E-Circular*, Vol. 71, pp. 39–43 (Washington, DC: Transportation Reserach Board).

LEE, M.S. and MCNALLY, M.G., 1998, Incorporating yellow-page databases in GIS-based transportation models. In *Proceedings of the American Society of Civil Engineers Conference on Transportation LandUse, and Air Quality*, S. Easa, (Ed.), pp. 652–661 (Reston VA: American Society of Civil Engineers).

LEIDNER, J.L., 2004a, Toponym resolution in text: 'Which Sheffield is it?'. In *Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR '04)*, M. Sanderson, K. Järvelin, J. Allan and P. Bruza (Eds), pp. 602 (Sheffield: Association for Computing Machinery).

LEIDNER, J.L., 2004b, Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference*, M. Sanderson, K. Järvelin, J. Allan and P. Bruza (Eds), pp. 33–40 (Sheffield: Association for Computing Machinery).

LEVINE, N. and KIM, K.E., 1998, The spatial location of motor vehicle accidents: A methodology for geocoding intersections. *Computers, Environment, and Urban Systems*, **22**(6), pp. 557–576.

LI, H., SRIHARI, R.K., NIU, C. and LI, W., 2002, Location normalization for information extraction. In *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7 (Morristown, NJ: Association for Computational Linguistics).

LIND, M., 2005, Addresses and address data play a key role in spatial infrastructure. Available online at: http://www.adresseprojekt.dk/files/ECGI Addr.pdf (accessed 29 January 2008).

LOCKYER, B., 2005, Opinion 04-1105, Legal opinion. Available online at: http://ag.ca.gov/opinions/pdfs/04-1105.pdf (accessed 29 January 2008).

MARK, D., SKUPIN, A. and SMITH, B., 2001, Features, objects, and other things: ontological distinctions in the geographic domain. In *COSIT 2001*, Vol. 2205, pp. 488–502 (Berlin: Springer-Verlag).

MARTINS, B., CHAVES, M. and SILVA, M., 2005a, Assigning geographical scopes to web pages. In *ECIR*, D.E. Losada and J.M. Fernández-Luna (Eds), Vol. 3408, pp. 564–567 (London: Springer).

MARTINS, B., SILVA, M. and CHAVES, M., 2005b, Challenges and resources for evaluating geographical IR. In *Proceedings of the 2005 ACM Workshop of Geographic Information Retrieval (GIR'05)*, C. Jones and R. Purves (Eds), pp. 17–24 (New York: Association for Computing Machinery).

MAYNARD, D., BONTCHEVA, K. and CUNNINGHAM, H., 2004, Automatic language-independent induction of gazetteer lists. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC '04)*, M.T. Lino, M.F. Xavier, F. Ferreira and R.S.R. Costa (Eds) (Lisbon: European Language Resources Association).

MCCURLEY, K.S., 2001, Geospatial mapping and navigation of the web. In *Proceedings of 10th International World Wide Web Conference*, V.Y. Shen and N. Saito (Eds), pp. 221–229 (New York: Association for Computing Machinery).

MCELROY, J.A., REMINGTON, P.L., TRENTHAM-DIETZ, A., ROBERT, S.A. and NEWCOMB, P.A., 2003, Geocoding addresses from a large population-based study: Lessons learned. *Epidemiology*, **14**(4), pp. 399–407.

MELISSA DATA CORPORATION, 2008, Data quality, data hygiene, and data cleansing, free lookups at Melissa data. Available online at: http://www.melissadata.com (accessed 29 January 2008).

MICHALOWSKI, M., THAKKAR, S. and KNOBLOCK, C.A., 2005, Automatically utilizing secondary sources to align information across sources. *AI Magazine, Special Issue on SemanticIntegration*, **26**(1), pp. 33–45.

MICROSOFT CORPORATION, 2008a, Live search. Available online at: http://local.live.com (accessed 29 January 2008).

MICROSOFT CORPORATION, 2008b, TerraServer-USA. Available online at: http://terraserverusa.com/ (accessed 29 January 2008).

MIKHEEV, A., MOENS, M. and GROVER, C., 1999, Named entity recognition without gazetteers. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, H.S. Thompson and A. Lascarides (Eds), pp. 1–8 (San Francisco, CA: Morgan Kaufmann Publishers).

MING, D., LUO, J., LI, J. and SHEN, Z., 2005, Features based parcel unit extraction from high resolution image. In *Proceedings of 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'05)*, W.M. Moon, (Ed.), Vol. 3, pp. 1875–1878 (Piscataway, NJ: Institute of Electrical and Electronics Engineers).

MORIMOTO, Y., AONO, M., HOULE, M.E. and MCCURLEY, K.S., 2003, Extracting spatial knowledge from the web. In *Proceedings of 2003 International Symposium on Applications and the Internet (SAINT 2003)*, pp. 326–333 (Piscataway, NJ: Institute of Electrical and Electronics Engineers).

NAVTEQ 2008, NAVTEQ Data. Available online at: http://www.navteq.com/about/data.html (accessed 29 January 2008).

OFFICE OF THE ASSESSOR, COUNTY OF LOS ANGELES, 2008, Property assessment information system. Available online at: http://assessormap.co.la.ca.us/mapping/viewer.asp (accessed 29 January 2008).

OPEN GEOSPATIAL CONSORTIUM, 2002, Gazetteer service profile of the web feature service implementation specification. OpenGIS® Project Document 02-076r3, R. Atkinson, and J. Fitzke (Eds).

PAULL, D., 2003, A geocoded national address file for Australia: The G-NAF what, why, who and when? (Grifith, ACT: PSMA Australia Limited). Available online at: http://www.addressonline.com.au/addressonline/home/GNAF What Why Who When.pdf (accessed 29 January 2008).

PEKAR, V. and EVANS, R., 2005, Automatic discovery of NLP resources on the web. In *Proceedings of 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC '05)*, P. Liddell, R. Siemens, A. Bia, M. Holmes, P. Baer, G. Newton and S. Arneil (Eds), pp. 161–163 (Victoria, BC: Humanities Computing and Media Centre, University of Victoria).

RATCLIFFE, J.H., 2001, On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science*, **15**(5), pp. 473–485.

REID, J., 2003, GeoXwalk: A gazetteer server and service for UK academia. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '03)*, T. Koch and I. Sølvberg (Eds), Vol. 2769, pp. 387–392 (London: Springer).

RIEKERT, W.F., 2002, Automated retrieval of information in the Internet by using Thesauri and gazetteers as knowledge sources. *Journal of Universal Computer Science*, **8**(6), pp. 581–590.

SAGARA, T., ARIKAWA, M. and SAKAUCHI, M., 2001, Web resource geographic location classification and detection. In *Proceedings of the Third International Conference on Information Integration and Web-based Applications & Services (iiWAS 2001)*, W. Winiwarter, S. Bressan and I. Ibrahim (Eds), pp. 399–409 (Linz: Austrian Computer Society).

SCHLIEDER, C., VÖGELE, T.J. and VISSER, U., 2001, Qualitative spatial representation for information retrieval by gazetteers. In *Proceedings of the 5th International Conference on Spatial Information Theory (COSIT 2001)*, D.R. Montello, (Ed.), Vol. 2205, pp. 336–351 (London: Springer).

SCHOCKAERT, S., COCK, M.D. and KERRE, E.E., 2005, Automatic acquisition of fuzzy footprints. In *Proceedings of Conference on the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops*, R. Meersman, Z. Tariand and P. Herrero (Eds), Vol. 3762, pp. 1077–1086 (London: Springer).

SINTICHAKIS, M. and CONSTANTOPOULOS, P., 1997, *A Method for Monolingual Thesauri Merging*, pp. 129–138 (New York: Association for Computing Machinery).

SMITH, B. and MARK, D., 1998, *Ontology and Geographic Kinds*, pp. 308–320 (Vancouver: International Geographical Union).

SMITH, D.A. and CRANE, G., 2001, Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '01)*, P. Constantopoulos and I. Sølvberg (Eds), Vol. 2163, pp. 127–136 (Darmstadt, London: Springer).

STAGE, D. and VON MEYER, N., 2005, An assessment of parcel data in the United States 2005 survey results. Technical report. Available online at: http://www.nationalcad.org/showdocs.asp?docid=170 (accessed 29 January 2008).

SURFLINE/WAVETRAK, INC., 2008, Surf report. Available online at: http://www.surfline.com/reports/report.cfm?id=4900 (accessed 29 January 2008).

TELE ATLAS, 2008, Our products – TeleAtlas.com, Available online at: http://www.teleatlas.com/OurProducts/MapData (accessed 29 January 2008).

UNITED STATES BUREAU OF THE CENSUS, 2008, Topographically integrated geographic encoding and referencing. Available online at: http://www.census.gov/geo/www/tiger (accessed 29 January 2008).

UNITED STATES POSTAL SERVICE, 2008a, Publication 28 – Postal addressing standards. Available online at: http://pe.usps.com/text/pub28/welcome.htm (accessed 29 January 2008).

UNITED STATES POSTAL SERVICE, 2008b, ZIP + 4 product file. Available online at: http://www.usps.com/ncsc/addressinfo/zip4.htm (accessed 29 January 2008).

URYUPINA, O., 2002, Extracting geographical knowledge from the Internet. In *Proceedings of the 2002 International Conference on Data Mining*, P. Constantopoulos and I. Sølvberg (Eds) (Piscataway, NJ: Institute of Electrical and Electronics Engineers).

URYUPINA, O., 2003, Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*, A. Kornai and B. Sundheim (Eds), pp. 81–89 (Edmonton, Alberta: Association for Computational Linguistics).

VAN RIJSBERG, C., 1979, *Information Retrieval* (London: Butterworth).

WANG, C., XIE, X., WANG, L., LU, Y. and MA, W.Y., 2005, Web resource geographic location classification and detection, pp. 1138–1139 (Chiba, (New York: Association for Computing Machinery).

WARD, M.H., NUCKOLS, J.R., GIGLIERANO, J., BONNER, M.R., WOLTER, C., AIROLA, M., MIX, W., COLT, J.S. and HARTGE, P., 2005, Positional accuracy of two methods of geocoding. *Epidemiology*, **16**(4), pp. 542–547.

WHITSEL, E.A., ROSE, K.M., WOOD, J.L., HENLEY, A.C., LIAO, D. and HEISS, G., 2004, Accuracy and repeatability of commercial geocoding. *American Journal of Epidemiology*, **160**(10), pp. 1023–1029.

WIECZOREK, J., GUO, Q. and HIJMANS, R.J., 2004, The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, **18**(8), pp. 745–767.

WILSON, J.P., LAM, C.S. and HOLMES-WONG, D.A., 2004, A new method for the specification of geographic footprints in digital gazetteers. *Cartography and Geographic Information Science*, **31**(4), pp. 195–203.

WITTEN, I.H., DON, K.J., DEWSNIP, M. and TABLAN, V., 2004, Text mining in a digital library. *International Journal on Digital Libraries*, **4**(1), pp. 56–59.

YAHOO!, INC., 2008, Yahoo! Maps, DRIVING DIrections, and Traffic. Available online at: http://maps.yahoo.com (accessed 29 January 2008).

YANG, D.H., BILAVER, L.M., HAYES, O. and GOERGE, R., 2004, Improving geocoding practices: Evaluation of geocoding tools. *Journal of Medical Systems*, **28**(4), pp. 361–370.

ZONG, W., WU, D., SUN, A., LIM, E.P. and GOH, D.H.L., 2005, On assigning place names to geography related web pages. In *Proceeding of the 2005 ACM/IEEE Joint Conference on Digital Libraries (JCDL 2005)*, M. Marlino, T. Sumner and F.M.S. III (Eds), pp. 354–362 (Washington, DC: Association for Computing Machinery).